Learning an Optimal Policy for Police Resource Allocation on Freeways

Brian Jackson, Taylor Howell, and Ola Shorinwa

Abstract—Anomalous detection is a well-studied area within computer science and artificial intelligence given its extremely useful applications in surveillance, driver assistance, and navigation. In this project, we apply a fundamental anomaly detection algorithm using Principal Component Analysis (PCA) and a Mixture of Gaussians model to the NGSIM highway dataset of real-world vehicle trajectories in congested traffic conditions. After identifying the feature distributions for anomalous drivers, we apply these distributions to a simulator to learn an optimal policy for police allocation to apprehend anomalous vehicles. We formulate the police allocation problem as a Markov Decision Process (MDP) and learn the optimal policy using reinforcement learning (RL). The optimal policy allocates police resources to apprehend drivers with higher citations, bypassing drivers with lower citations.

Index Terms—EM algorithm, Principal Component Analysis, Markov decision process, machine learning, reinforcement learning

I. INTRODUCTION

EACH year over 32,000 driving-related deaths and 2 million injuries occur on US roads [1]. We posit that automatic traffic surveillance and efficient allocation of law enforcement resources to apprehend anomalous drivers can improve road safety. The purpose of this project is to develop a framework for identifying "anomalous" drivers on freeways and optimally allocating police resources to apprehend and cite these drivers. In the context of this project, anomalous drivers may or may not be breaking laws such as speeding, tailgating, or weaving in and out of traffic; however, if their driving style is significantly different than the rest of the cars around them, some benefit may be derived in identifying the underlying cause of their observed behavior. For instance, it may be possible to identify drunk or impaired drivers, distracted drivers, and vehicle thieves or kidnappers. Surveillance systems that automatically detect anomalous drivers can act as useful tools to help law enforcement agencies more efficiently sift through the thousands of drivers on the road and focus on ones that "stand out."

Once these drivers are identified, the allocation of scarce resources (in this case, highway patrol vehicles) to optimally apprehend suspicious or unlawful drivers is yet another challenge. In this project, we aim to solve both of these problems, namely: identification of anomalous drivers, and finding an optimal policy for allocating police resources. The first problem is solved by learning a mixture of Gaussians model from a data set taken from actual highway data. The second problem

Department of Mechanical Engineering, Stanford University

is framed as a Markov decision process (MDP) and solved using reinforcement learning (RL) techniques.

The problem of detecting anomalous driver is a well-studied problem given its valuable applications in surveillance, crisis prevention, driver assistance, and navigation. [2] gives an overview of various approaches to anomalous behavior detection, which are categorized into the following classes: classification-based, parametric or non-parametric statistical, nearest neighbor-based, clustering-based, spectral techniques, and information theoretic. Many studies have taken a similar approach to the one used in the current project. Notably, Morris and Trivedi [3] propose an algorithm for extracting trajectory information from camera data, compressing the feature subspace using Principal Component Analysis (also used in the current project), and then classifying the data using a variety of unsupervised learning techniques including K-means, fuzzy-C-means, and neural networks. New trajectories are assigned a score rating on how anomalous they are, based upon the distance to the nearest cluster. Other approaches have included support-vector machines (SVMs) [4], Hidden-Markov Models clustered using K-means [5], and semi-supervised learning [6]. The field of anomaly detection is a broad and very active area of research, and presenting any sort of in-depth analysis of the various approaches is beyond the scope of this project. However, based on the cited papers, SVM techniques are advantageous given their computational efficiency, while the more traditional approach is advantageous given that it builds a distribution that models "typical" driver behavior. The semi-supervised learning approach showed promising results but requires human-in-the-loop training and is susceptible to undetected anomalies. It seems current state-of-the-art techniques like the one presented by An and Cho [7] use deep-learning-based autoencoders, that reduce feature dimensionality and increase abstract generalization. An and Cho present an algorithm for variation autoencoders (VAE) that combine deep autoencoders with variational inference. Since it is a generative model, it can also be used to derive the underlying model and analyze the underlying cause of the anomaly.

Resource allocation is a fairly well-studied field, and the topic of police allocation in particular has received some attention. However, applying state-of-the-art decision making algorithms to police allocation is very limited [8], and typically based on optimizing deployment location rather than deciding whether or not to deploy limited resources. Our project therefore seems to be a novel application of decision-making processes to learning an optimal policy for deploying police in an uncertain environment with variable rewards.

II. METHODOLOGY

The project is divided into three principal components: 1) preprocessing: analyzing and compressing real-world highway-driving data, 2) anomaly detection: identifying anomalous drivers from the dataset, and 3) the MDP: learning an optimal policy for deploying police resources to cite anomalous drivers and maximize ticket revenue. We assume that it's possible to detect anomalous driving behavior that corresponds with traffic offenses on freeways and that it's possible to indirectly make the freeway safer by maximizing ticket revenue. We will model highway driving behavior as a mixture of Gaussians and characterize atypical driving behavior as lying outside the modeled distributions. We will be working with the NGSIM US Highway 101 data set, which tracks vehicle trajectories in congested morning traffic along a half mile section of Highway 101. We will use the EM algorithm to train the model with features from these driving trajectories.

Data Processing

We are using real-world driver data from the NGSIM US-101 dataset. The data was captured using a set of cameras set atop a 36-story building. This data has been processed into vehicle trajectories that track individual vehicles and report features such as width, length, velocity, position, lane, etc. However, the original trajectory data contains significant noise in the position, velocity, and acceleration estimates. We are using data that has been post-processed by members of Stanford Intelligent Systems Laboratory (SISL), which improves state estimates using a Kalman filter. The resulting data consists of 45 minutes of vehicle trajectories, including about 6000 unique vehicle trajectories. From this filtered data, we have constructed a training set consisting of features averaged over 15-second time intervals with 5 second spacing, resulting in about 64,000 unique driver examples.

Each of these examples contains six features: average velocity, average acceleration, maximum velocity, number of lane changes, average deviation from lane centers, and standard deviation from lane centers. The velocity, acceleration, and deviation from lane centers were all given by the filtered dataset. The averages, maximums and standard deviation were taken over the 15-second time window of interest. The number of lane changes was determined by counting the number of times the lane deviation varied by more than a fixed threshold in a given time step (in either direction). The resulting features should give a decent depiction of driver behavior over a moderate time window and be sufficient to identify anomalous behavior.

Anomaly Detection

1) EM Algorithm: We assumed that each example came from a joint distribution $p(x, z) = p(x|z)p(z)$ consisting of four Gaussians (the number of latent distributions was selected by balancing the number of anomalous drivers we wanted to consider in the police allocation problem). Thus, each $x^{(i)}$ was obtained by randomly choosing $z^{(i)} \in \{1, 2, ..., k\}$

and then sampling $x^{(i)}$ from the corresponding Gaussian distribution. We intend to learn the parameters of the joint distribution $p(x, z; \theta)$ for the training examples where $z^{(i)}$ represents the latent variables and $\theta = \{\phi, \mu, \Sigma\}$. Finding a closed form solution for the maximum likelihood estimates of θ is difficult; hence, we used the EM algorithm for estimating θ. The EM algorithm works by constructing a lower bound on the log-likelihood of the training set $\ell(\theta)$ before optimizing the lower bound. The algorithm is given below.

E-step: For each i,

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta)$$

M-step: Set

$$\theta := \operatorname*{argmax}_{\theta} \sum_{i}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

We use the estimated parameters in fitting a mixture of Gaussians to the NGSIM data set and identify outliers using a threshold $p(x) < 0.0025$.

2) Principal Component Analysis: We used principal component analysis as a dimensionality reduction algorithm to map the extracted features from 6 features to different feature spaces ranging from sizes 2 to 5. The first k principal components of the data were selected to maximize the variance of the projections where the first principal component satisfies the optimization problem:

$$u := \operatorname*{argmax}_{u:u^T u = 1} \frac{1}{m} \sum_{i=1}^{m} (x^{(i)^T} u)^2$$

We apply the EM algorithm to each of the resulting feature spaces to obtain a model for classifying anomalous drivers. The features of the anomalous drivers flagged by each model was analyzed to evaluate the performance of each model.

The MDP

We have formulated the task of allocating police resources on freeways as a discounted infinite-horizon Markov decision process.

MDP Formulation: The state is defined by two variables: the police state and the driver state. The police state is the vector $P = [p_1 \ldots p_{n_p}] \in \mathbb{Z}^{n_p}$ of police car states $p \in \{0, 1, ..., p_{max}\}$. Each integer p_i represents the number of time steps until the police car i is again available for allocation. $p_i = 0$ indicates that police i is ready for allocation. After being allocated, $p_i = p_{max}$ and decrements by one each time step. The driver state is the vector $D = [d_1 \ldots d_{n_d}] \in \mathbb{Z}^{n_d}$ of driver states $d \in \{1, ..., d_{max}\}$. The integer values d_j correspond to the state of the anomalous vehicle j, defined as by Table I.

The size of the state space is $(p_{max})^{n_p} \times (d_{max})^{n_d}$. For our final solution, we set $n_p = 1, n_d = 2,, p_{max} = 5$, and $d_{max} = 4$. ($n_p = 2, n_d = 4,, p_{max} = 5$, and $d_{max} = 4$).

The action space defines the allocation of police cars in P. The action $a = \{0, 1, \ldots, n_p\} \in \mathbb{Z}$ corresponds to the number of citations to deliver. It is assumed that the vehicles in D with the highest reward will be cited.

d	Reward	Description	Conditions
1	50	No citation	Not 2 or 3
2	100	Speeder	Speed > 17 m/s
3	150	Weaver	# lane changes > 3
4	250	Both	Both 2 & 3

TABLE I

DEFINITION OF ANOMALOUS VEHICLE STATES

The rewards are deterministically assigned: each vehicle state p_j is assigned a particular reward according to the simulator which queries the anomaly detector, as shown in Table I. If $a > 0$, the resulting reward is the sum of the a highest rewards corresponding to states in D. If $a > 0$ and $p_i \neq 0, \forall p_i \in P$ (no police available), then a penalty of -100 was assigned.

Scene simulation: Since the NGSIM data set only gives us a finite amount of experience, we created a simulator that samples from distributions based upon the NGSIM data set. The simulator maintains an internal "scene" of anomalous vehicles. A scene contains 10 vehicles, specified by the features listed in section II, in addition to a unique vehicle ID and a time stamp. The time stamp of each vehicle is decremented at each time step and vehicles with negative time stamps are removed from the scene. The features are sampled from independent distributions. The authors acknowledge that in reality the anomalous drivers are more likely to belong to a multivariate distribution that encodes dependence between different features, such as average velocity and number of lane changes, but for simplicity these features were assumed to be independent, and the resulting samples will consequently not be guaranteed to be anomalous. However, only n_d of the 10 vehicles are selected from each scene, according to highest reward. This is a fair assumption since it would clearly be sub-optimal to issue a citation to a vehicle known to yield less reward (given our previous assumption that the rewards are deterministic). At each time step, new vehicles are added to keep the total number of vehicles in each scene constant. New vehicles were always initialized with a time stamp of p_{max} (described above). The time stamp was included with the purpose of modeling the fact that drivers can only be observed for a fixed amount of time and that the police only have a finite amount of time to decide whether or not to apprehend the vehicle. After adding the new vehicles, they are assigned a state value as a deterministic function of their features (maximum velocity and number of lane changes), according to the conditions listed in Table I. The distribution of vehicle state values used in the simulation is shown in Figure 1. It should be noted that the threshold for speeding (17 m/s, or 38 mph) is lower than the actual speed limit (65 mph). The average maximum speed of the drivers in the data set was found to be 11.176 m/s which was expected as the data was recorded during heavily congested morning traffic. Thus, the chosen speed limit reflects anomalous behavior among the population of drivers in the data set. Figure 1) shows the distribution over the features of the anomalous drivers.

At each time step, the simulator accepts a state/action tuple and returns the new state and reward. Based on the input state and action, it issues citations to the vehicles in D with
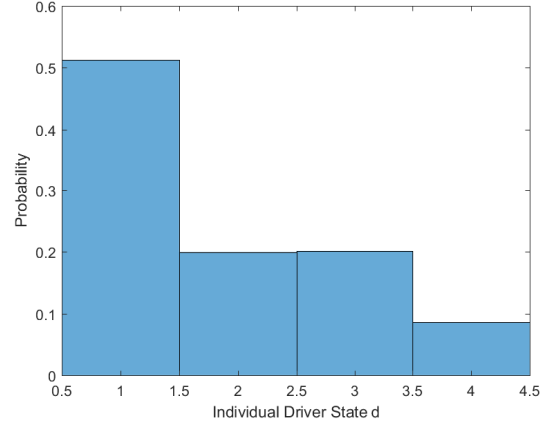

Fig. 1. Distribution of individual vehicle states used by simulation

the highest reward and removes them from the scene. It then updates the scene by decrementing the time stamps and adding new vehicles.

Solving the MDP: We use model-based reinforcement learning to find an optimal policy for the MDP. The agent gains experience by interacting with the simulator for $n = 100$ time steps before updating the transition $T(s'|s, a)$ and reward $R(s, a)$ models using maximum likelihood estimates. Asynchronous value iteration is performed on the Bellman equation (Eq. 1).

$$U^*(s) := \max_a \left(R(s, a) + \gamma \sum_{s'} T(s'|s, a)U^*(s') \right) \quad (1)$$

An update to the value function was assumed to have converged when the maximum absolute difference in value functions $||U_k - U_{k-1}||_\infty < 0.005$. The learning procedure was repeated until 10 consecutive updates of the the value function converged on the first iteration. The optimal policy $\pi^*(s)$ was obtained from the optimal value function.

$$\pi^*(s) := \operatorname*{argmax}_a \left(R(s, a) + \gamma \sum_{s'} T(s'|s, a)U^*(s') \right)$$

To evaluate the optimal policy, we calculate the cumulative rewards per allocation over a 1000 step episode (and average the results over 5 episodes) for the optimal policy and similarly compare it against an randomly-send policy that randomly assigns police vehicles to apprehend offending drivers if police are available and an always-send policy that assigns the maximum number of available police vehicles at each time step.

III. RESULTS

A. Anomaly Detection

New feature sets of sizes ranging from 2 to 4 were constructed from the highway data set using PCA. These feature sets were evaluated against the baseline model consisting of six features. The anomaly detection model was created by fitting four Gaussians to each feature set. Anomalous drivers were

Size of feature space	Percentage of anomalous drivers (%)
2	0.249
3	3.648
4	5.356
5	29.514

TABLE II

PERCENTAGE OF ANOMALOUS DRIVERS FOR EACH MODEL
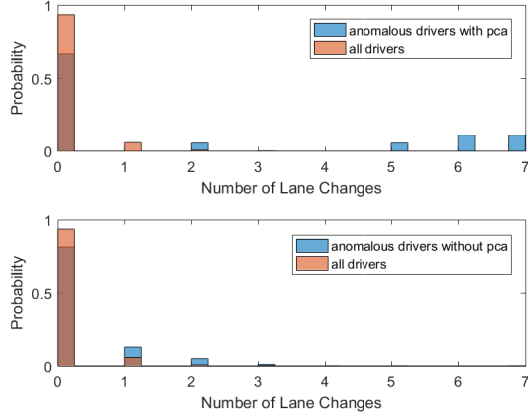
Fig. 2. Histogram of number of lane changes of anomalous drivers for the two-feature and baseline models
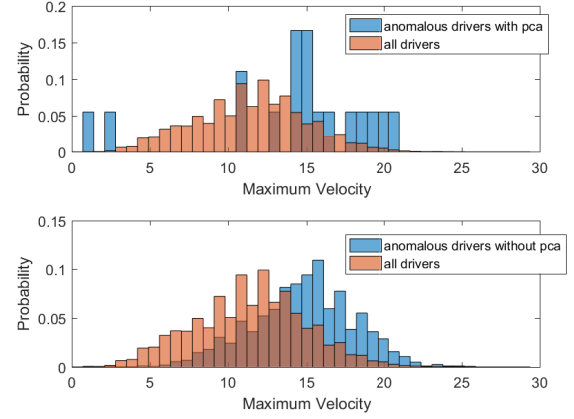

Fig. 3. Histogram of the maximum velocity of anomalous drivers for the two-feature and baseline models

Fig. 4. Histogram of average acceleration of anomalous drivers for the two-feature and baseline models

identified using a threshold $p(x) < 0.0025$. Table II shows the percentage of anomalous drivers flagged by each model. The number of anomalous drivers decreases with the size of the feature space. The baseline model flagged about 30% of drivers as atypical while the two-feature set model identified a substantially lower percentage of atypical drivers, 0.25%. The three-feature and four-feature models classified about 4% and 6% of drivers as atypical respectively.

The features of the flagged drivers were analyzed to further evaluate the effects of dimensionality reduction using PCA. Figure 2 shows the histogram of the number of lane changes of outliers identified by the two-feature and baseline models. The two-feature model flags drivers having a high number of lane changes with high probability compared to the baseline model. The drivers with zero lane changes which were classified as anomalous exceeded the acceptable limits on other features.

Figure 3 shows the normalized histograms of the maximum velocity of the anomalous drivers identified by the two-feature and baseline models. The two-feature model performs remarkably well in identifying drivers moving at high velocities in addition to slow-moving drivers who might have been distracted. The baseline model seems more susceptible to noise, flagging a greater proportion of drivers with acceptable maximum velocities as anomalous.

Figure 4 shows the normalized histograms of the average acceleration of anomalous drivers identified by the two-feature and baseline models. The two-feature model performs considerably well in identifying drivers with extreme accelerations, flagging such drivers as anomalous with a high probability. In contrast, drivers with acceptable accelerations have a higher probability of being flagged anomalous by the baseline model compared with drivers with extreme accelerations. The results show that the application of PCA helped reduce the suscepti-

bility of the model to noisy observations, ultimately improving the accuracy of the anomaly detection algorithm. The PCA model retained the valuable variations among the feature attributes while discarding noisy variations which resulted in the model flagging a lower percentage of drivers as anomalous. The application of PCA also provided computational benefits by reducing the size of the feature sets. Such computational benefits scale with the size of the training set.
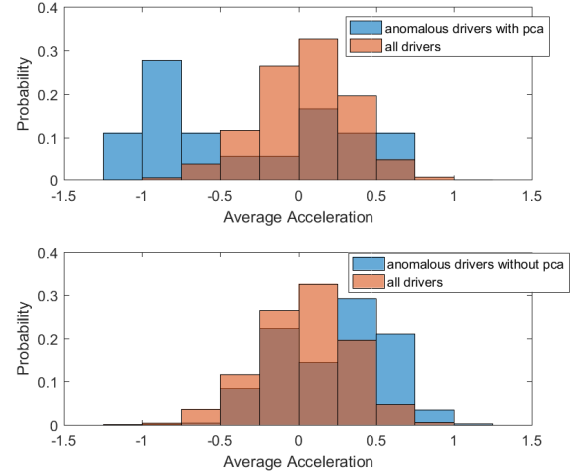
The simulator for the MDP generates new anomalous vehicles by sampling from independent distributions over each feature. The distributions over these features were created from the anomalous drivers identified by the two-feature model. The normalized histograms of the features are multi-modal; hence, the features could not be modeled as a unimodal Gaussian distribution. We estimated the distribution using a normal kernel distribution with different bandwidths. The bandwidth parameter controls the smoothness of the probability density curve. Figure 5 shows the fitted probability distributions over the features for different bandwidths. The default bandwidth parameter shown in the figure results from optimizing the
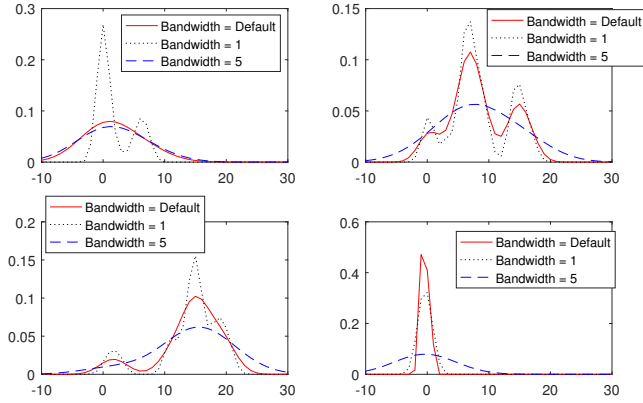
Fig. 5. Kernel distribution over the (a) number of lane changes, (b) average velocity, and (c) maximum velocity and (d) average acceleration for different bandwidths

State			Value	Policy		
d_1	d_2	p_1	U*	Optimal	Always	Random
1	1	0	333.209	0	1	0
2	1	0	341.488	0	1	0
3	1	0	372.812	1	1	1
4	1	0	454.712	1	1	1
2	2	0	341.113	0	1	0
3	2	0	377.063	1	1	1
4	2	0	475.914	1	1	0
3	3	0	377.997	1	1	1
4	3	0	478.403	1	1	0
4	4	0	496.187	1	1	1

TABLE III
POLICIES FOR CRITICAL STATES

mean integrated square error. A bandwidth parameter of one provided better estimations of the probability density curves of the number of lane changes, average velocity, and maximum velocity, showing the major peaks in the probability density curve for these features. The kernel distribution using the theoretically optimal bandwidth parameter provided a better estimate of the probability density curve of the average acceleration and was used by the simulator.

Fig. 6. Learning Curve

B. Optimal Policy

For the simple MDP with one police and two anomalous drivers, solving for the optimal policy took 2135.18 seconds, converging after 21,000 epochs (100 steps of experience per epoch). An ϵ-greedy strategy was used with $\epsilon = 0.1$ and a discount factor $\gamma = 0.9$. For any state with no police to allocate (i.e., $\sum_{i=1}^{n_p} p_i = 0$) the police learned to do nothing (i.e., action = 0) which is optimal since taking any other action in this state will result in a negative reward. To better understand the policy, we examine the policy at critical states where police are available for allocation (i.e., $p_i = 0$). For these states, we find that there is some threshold such that drivers with low reward values are not cited. We found that drivers with $d_i < 3$ were not cited. This demonstrates that our agent was able to learn an intelligent policy that enables it to delay immediate reward for greater future reward. We compare the optimal policy with a randomly-send policy (if $\sum_{i=1}^{n_p} p_i > 0$ then a random number of available police are allocated, else no police are allocated) and always-send policy (number of police allocated = $\sum_{i=1}^{n_p}$). Because we are interested in efficiently allocating resources, we compare the average cumulative reward per allocation for each policy for a 1000 step episode for 5 trials and then average the results. The results of the learned optimal policy for the during training versus the "expert" policies are shown in Figure 6.
The actual learned policy (and value function $U*$) for key states is shown in Table III against the expert policies.
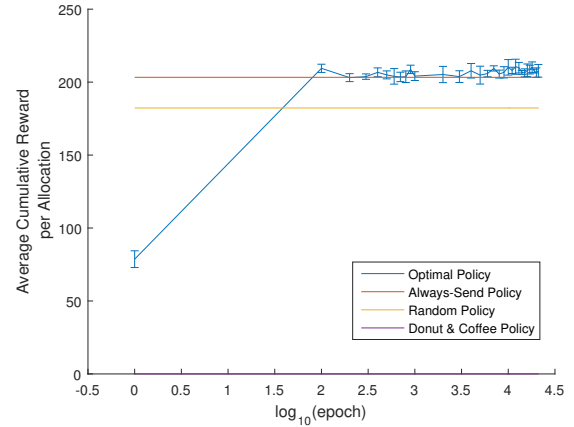
IV. CONCLUSION

Driving behavior on freeways can be modeled by fitting a mixture of Gaussians to the trajectory data of the drivers. Anomalous drivers can be identified using the resulting model with a given threshold $p(x) < \epsilon$, indicating the probability of a driver's behavior lying outside the acceptable range. We applied PCA as a dimension-reduction algorithm before fitting the model. Our results indicate that the application of PCA improved the accuracy of the algorithm. The PCA models also reduced the number of flagged drivers to more realistic proportions between 0.2% and 0.5%. Such findings suggest that PCA improved the resilience of our model to sensor and driver noise. We were also able to learn an optimal policy for police allocation. As expected, the optimal policy does not allocate police vehicles to anomalous drivers with low citation rewards, preferring to reserve those vehicles for future time steps when drivers with higher citations might be identified. Consequently, our agent was able to maximize the citation rewards with a minimal number of police allocations.

The features of anomalous drivers are not independent as assumed in our generation of new anomalous drivers. To address this, we intend to train a generative adversarial network to generate new anomalous drivers for the MDP simulator. We also expect the optimal policy (reserving police vehicles for drivers with higher citations) to scale to a larger state space with more police vehicles and anomalous drivers and plan to test this hypothesis.

ACKNOWLEDGMENT

REFERENCES

[1] E. K. Sauber-Schatz, "Vital signs: motor vehicle injury preventionunited states and 19 comparison countries," *MMWR. Morbidity and mortality weekly report*, vol. 65, 2016.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[3] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE transactions on circuits and systems for video technology*, vol. 18, no. 8, pp. 1114–1127, 2008.

[4] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Transactions on Circuits and Systems for video Technology*, vol. 18, no. 11, pp. 1544–1554, 2008.

[5] N. Suzuki, K. Hirasawa, K. Tanaka, Y. Kobayashi, Y. Sato, and Y. Fujino, "Learning motion patterns and anomaly detection by human trajectory analysis," in *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*. IEEE, 2007, pp. 498–503.

[6] R. R. Sillito and R. B. Fisher, "Semi-supervised learning for anomalous trajectory detection." in *BMVC*, vol. 27, 2008, pp. 1025–1044.

[7] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," Technical Report, Tech. Rep., 2015.

[8] N. Adler, A. S. Hakkert, J. Kornbluth, and M. Sher, "Location-allocation models for traffic police patrol vehicles on an interurban network," *Springer*, pp. 9–31, 2014.